EDUCATION**SECTOR** REPORTS

January 2008

RUSH TO JUDGMENT:

Teacher Evaluation in Public Education

By Thomas Toch and Robert Rothman

TABLE OF CONTENTS

Cover photo courtesy of Cindy England/iStockphoto.

1201 Connecticut Ave., N.W., Suite 850, Washington, D.C. 20036
202.552.2840 • www.educationsector.org

ACKNOWLEDGEMENTS

This report was funded in part by KnowledgeWorks Foundation and the William T. Grant Foundation. We thank the foundations for their support but acknowledge that the findings and conclusions presented in the report are those of the authors alone and do not necessarily represent the opinions of the foundations.

Danny Rosenthal, Claire Williams, and Troy Scott supplied research support for this report. And we thank the many teachers, administrators, researchers, and other experts that we spoke with in preparing the report.

ABOUT THE AUTHORS

THOMAS TOCH is co-founder and co-director of Education Sector. He is on the board of advisors of DC Preparatory Academy. He can be reached at ttoch@educationsector.org.

ROBERT ROTHMAN is a principal associate for the Annenberg Institute for School Reform at Brown University. He can be reached at robert_rothman@brown.edu.

ABOUT EDUCATION SECTOR

Education Sector is an independent education policy think tank devoted to developing innovative solutions to the nation's most pressing educational problems. We are nonprofit and nonpartisan, both a dependable source of sound thinking on policy and an honest broker of evidence in key education debates throughout the United States.

Generations of education reformers have sought to strengthen the ranks of public school teaching. And, almost always, their recommendations have included abolishing what is known as the single salary schedule, the nearly universal practice in public education of paying teachers not on the basis of performance, but strictly on the basis of the college credits they've amassed and the years they've taught.

The organizers of a 1955 White House education conference counseled in a report to President Eisenhower that, "Every effort must be made to devise ways to reward teachers according to their ability without opening the school door to unfair personnel practices." The authors of "A Nation at Risk," the stinging 1983 indictment of public education, had the same message: Teacher salaries needed to be "professionally competitive, market-sensitive, and performance-based."[1]

So did the Teaching Commission, a 19-member panel of national luminaries chaired by Louis Gerstner, the former chairman of IBM. "By precluding the possibility of performance-driven compensation, we fail to attract more talented and motivated individuals to our schools," it warned in 2004.[2]

But though there have been many performance-pay experiments in public education since the advent of the single salary schedule back in the 1920s, most haven't lasted more than a couple of years.[3] That shouldn't be a surprise, despite performance pay's many influential advocates. Teachers unions are partly responsible; many of them have fought performance pay aggressively since their rise to power in the 1960s. But there's another, rarely mentioned reason why performance pay has never caught on in public education: Rewarding teachers on the basis of their performance requires a credible system of measuring the quality of teachers' work—something that the vast majority of public schools don't have.

A host of factors—a lack of accountability for school performance, staffing practices that strip school systems of incentives to take teacher evaluation seriously, union ambivalence, and public education's practice of using teacher credentials as a proxy for teacher quality—have resulted in teacher evaluation systems throughout public education that are superficial, capricious, and often don't even directly address the quality of instruction, much less measure students' learning.

The troubled state of teacher evaluation is a glaring and largely neglected problem in public education, one with consequences that extend far beyond the performance-pay debate. Because teacher evaluations are at the center of the educational enterprise—the quality of teaching in the nation's classrooms—they are a potentially powerful lever of teacher and school improvement. But that potential is being squandered throughout public education, an enterprise that spends $400 billion annually on salaries and benefits.

The task of building better evaluation systems is as difficult as it is important. Many hurdles stand in the way of rating teachers fairly on the basis of their students' achievement, the solution favored by many education experts today. And it's increasingly clear that it's not enough merely to create more-defensible systems for rewarding or removing teachers. Teacher evaluations pay much larger dividends when they also play a role in improving teaching.

This report explores the causes and consequences of the crisis in teacher evaluation. And it examines a number of national, state, and local evaluation systems that point to a way out of the evaluation morass. Together, they demonstrate that it's possible to evaluate teachers in much more productive ways than most public schools do today.

Drive-Bys

It's hard to expect people to make a task a priority when the system they are working in signals that the task is unimportant. That's the case with teacher evaluation.

Public education defines teacher quality largely in terms of the credentials that teachers have earned, rather than on the basis of the quality of the work they do in their classrooms or the results their students achieve.

There's logic to having reading teachers enter classrooms knowing how kids learn to read and in having algebra teachers armed with strategies for teaching quadratic equations—the sorts of skills that are supposed to be reflected in teaching credentials. But recent studies have found that such qualifications don't guarantee effective teachers. A 2005 report on 9,400 Los Angeles teachers by Thomas Kane of Harvard and Douglas Staiger of Dartmouth, for example, found no meaningful difference in the achievement results of students taught by teachers who were certified and those taught by teachers who lacked certification. In some instances, the unlicensed teachers produced substantially higher results than their certified counterparts.[4]

In its pursuit of school improvement, the federal No Child Left Behind Act (NCLB) has unwittingly intensified public education's culture of credentialism. The law has sought to improve teacher quality by requiring that schools employ only "highly qualified" teachers. But it mandates that states use the qualifications that teachers bring to the classroom—rather than their performance as teachers—as the measure of whether teachers meet the law's standard.

The single salary schedule, a product of the sexism and favoritism that plagued the teaching profession at the beginning of the twentieth century, both reflects and reinforces public education's emphasis on credentials at the expense of performance. Paying teachers with the same credentials—and the same years of experience—exactly the same salaries devalues the importance of their effectiveness in the classroom and diminishes the significance of teacher evaluations.[5]

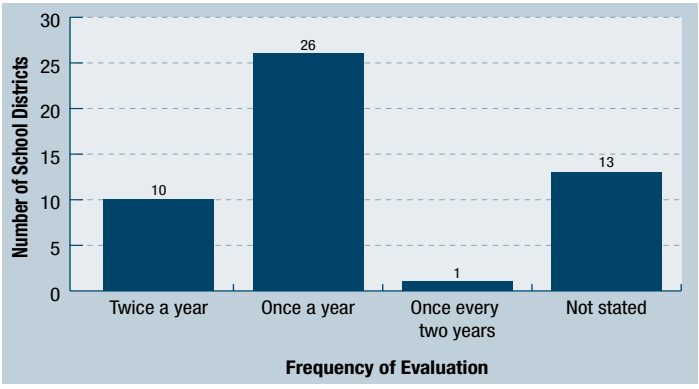
It's not surprising, then, that measuring how well teachers teach is a low priority in many states. The nonprofit National Council on Teacher Quality (NCTQ) reports that, despite many calls for performance pay coming from state capitals, only 14 states require school systems to evaluate their public school teachers at least once a year, while some are much more lax than that. Tennessee, for example, requires evaluations of tenured teachers only twice a decade.[6]

An NCTQ analysis of the teacher contracts in the nation's 50 largest districts (which enroll 17 percent of the nation's students) suggest that not much teacher evaluation is enshrined in local regulations, either. Teachers union contracts dictate the professional requirements for teachers in most school districts. But the NCTQ study found that only two-thirds of them require teachers to be evaluated at least once a year and a quarter of them require evaluations only every three years.[7]

The evaluations themselves are typically of little value—a single, fleeting classroom visit by a principal or other building administrator untrained in evaluation wielding a checklist of classroom conditions and teacher behaviors that often don't even focus directly on the quality of teacher instruction. "It's typically a couple of dozen items on a list: 'Is presentably dressed,' 'Starts on time,' 'Room is safe,' 'The lesson occupies students,'" says Michigan State University Professor Mary Kennedy, author of *Inside Teaching: How Classroom Life Undermines Reform*, who has studied teacher evaluation extensively. "In most instances, it's nothing more than marking 'satisfactory' or 'unsatisfactory.'"

It's easy for teachers to earn high marks under these capricious rating systems, often called "drive-bys," regardless of whether their students learn. Raymond Pecheone, co-director of the School Redesign Network at Stanford University and an expert on teacher evaluation, suggests by way of example that a teacher might get a "satisfactory" check under "using visuals" by hanging

Figure 1. Evaluation Requirements of Untenured Teachers in the 50 Largest U.S. School Systems*



*As required by collective-bargaining contracts.

Source: Education Sector analysis of data from the National Council on Teacher Quality, "Teacher Rules, Roles and Rights," http://www.nctq.org/cb/.

up a mobile of the planets in the Earth's solar system, even though students could walk out of the class with no knowledge of the sun's role in the solar system or other key concepts. These simplistic evaluation systems also fail to be remotely sensitive to the challenges of teaching different subjects and different grade levels, adds Pecheone.

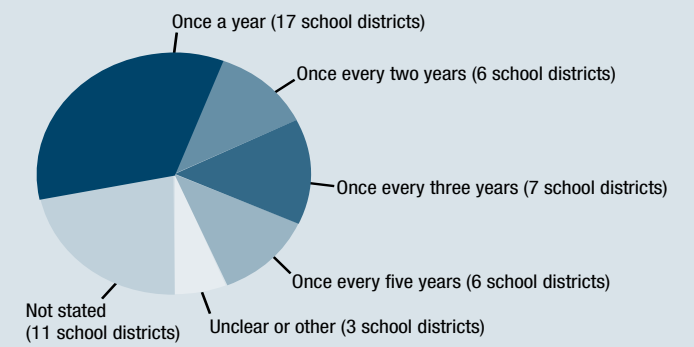
Unsurprisingly, the results of such evaluations are often dubious. Donald Medley of the University of Virginia and Homer Coker of Georgia State University reported in a comprehensive 1987 study titled "The Accuracy of Principals' Judgments of Teacher Performance" that the research up to that point found the relationship between the average principal's ratings of teacher performance and achievement by the teachers' students to be "near zero."[8]

Principals fared better in a recent study by Brian Jacob of Harvard's Kennedy School of Government and Lars Lefgren of Brigham Young University that compared teacher ratings to student gains on standardized tests. Principals were able to identify with some accuracy their best and worst teachers—the top 10 or so percent and the bottom 10 or so percent—when asked to rate their teachers' ability to raise math and reading scores.[9]

But principals don't put even those minimal talents to use in most public school systems. A recent study of the Chicago school system by the nonprofit New Teacher Project, for example, found that 87 percent of the city's 600 schools did not issue a single "unsatisfactory" teacher rating between 2003 and 2006. Among that group of schools were 69 that the city declared to be failing educationally. Of all the teacher evaluations conducted during those years, only .3 percent produced "unsatisfactory" ratings, while 93 percent of the city's 25,000 teachers received top ratings of "excellent" or "superior."[10]

And principals use evaluations to help teachers improve their performance as rarely as they give unsatisfactory ratings. They frequently don't even bother to discuss the results of their evaluations with teachers. "Principals are falling prey to fulfilling the letter of the law," says Dick Flannery, director of professional development for the National Association of Secondary School Principals, a principals' membership organization. "They are missing the opportunity to use the process as a tool to improve instruction and student achievement."

Figure 2. Evaluation Requirements of Tenured Teachers in the 50 Largest U.S. School Systems*



*As required by collective-bargaining contracts.

Source: Education Sector analysis of data from the National Council on Teacher Quality, "Teacher Rules, Roles and Rights," http:/www.nctq.org/cb/.

Test Scores

One school of reformers, including many of today's generation of performance-pay advocates, would evaluate teachers on the basis of their students' achievement. It's a reasonable strategy: It's the most direct way to measure teacher performance, and teaching is ultimately about helping students learn. But currently the only way to measure student achievement on a large scale is with standardized test scores. And that makes the student-achievement strategy difficult.

For one thing, only about half the nation's teachers teach subjects that are tested. It wouldn't be possible to use student test scores in individual teacher evaluations for the other half.

Secondly, a majority of the standardized tests that would be used in teacher evaluations today—statewide tests required by NCLB—focus on low-level skills such as the recall or restatement of information and on only a few subjects, primarily reading, math, and science. They don't measure more advanced skills such as expository writing or an ability to think creatively or analytically, and they sidestep history, art, music, and other subjects. As a result, they can't capture a teacher's skill in energizing students to learn astronomy or in scaffolding a series of lessons that draw students into the life of a novel. "They privilege very low level pedagogy," says Pecheone. "The best teachers, those that have a wider teaching repertoire and are able to engage students beyond the basics, are at a disadvantage." It might be reasonable, as a result,

to use test scores as a factor in weeding out the weakest teachers, but they wouldn't be as good at identifying the best teachers.

It's also the case that teachers are dealt different hands from classroom to classroom and school to school. Some work with students who have privileged backgrounds, who previously have had very good teachers, or who are very bright. Others teach students who are less fortunate, less well-prepared, and less capable. So handing teachers high ratings merely if their students have high test scores would result in many misjudgments of teachers' true abilities.

Teachers of rich kids may do a lousy job in the classroom, but their students nonetheless may get higher test scores than their less-privileged peers. And teachers of less-privileged students may do a great job, only to have their students come up short compared to students with more advantages. Evaluation systems with this unfairness built into them would create a strong incentive for teachers to abandon challenging students and the schools that enroll them.

The most common way of reporting student performance on standardized tests is as a percentage of teachers' students scoring high enough to meet state standards. But states have wildly varying standards. So it would be far easier for teachers to earn satisfactory ratings under an evaluation system using test scores in states with low standards.

Consider the cases of Colorado and South Carolina. The Thomas B. Fordham Institute and the Northwest Evaluation Association, a testing company, recently calculated how students scoring just high enough to meet their state standards in reading and math would do on a national test. They found that Colorado's eighth-graders would score at the 14th percentile on the national test in reading, and their counterparts in South Carolina, where the reading standards are much higher, would score at the 71st percentile.[11]

It's not surprising, as a result, that many teachers are strongly opposed to evaluations based substantially or exclusively on student test results. So there's an added risk to such evaluations: The people who would be subjected to them don't think they're credible.

The news isn't all bleak on the testing front: Solutions are emerging to the unfairness of testing-based evaluations to teachers who work with disadvantaged and under-prepared students. By calculating a teacher's performance on the basis of how much their students' test scores increase over the course of a school year, these solutions are able to isolate the effects of individual teachers on student learning and determine the "value added" they provide. Teachers don't get rewarded for having a class of high achievers or penalized for teaching low-performers.

There are two catches, though. First, the very small numbers of students that some teachers have makes it difficult to calculate with statistical confidence their impact on their students' test scores. Second, only about a third of the states currently have the computer systems necessary to link teachers and students in ways needed to do value-added calculations.[12] Some school systems contract directly with private companies to do the calculations. But the numbers are tiny, only about 300 out of 14,000 nationally.[13]

A New Model

A small number of local, state, and national initiatives have sought a different solution to drive-by evaluations—comprehensive evaluation systems that measure teachers' instruction in ways that promote improvement in teaching.

Charlotte Danielson has had an important role in the emergence of the comprehensive systems. In the early 1990s, Danielson, now 65, was working at the Educational Testing Service (ETS), the Princeton, N.J.–based testing company best known for its college-admissions exams such as the SAT and the GRE, when she joined a team developing a package of teacher-licensure examinations, known as Praxis. Praxis I is a basic reading and math test. Praxis II is a series of tests of subject-matter and teaching knowledge.

Danielson worked on Praxis III, which sought to measure the classroom skills of neophyte teachers. ETS-trained evaluators were to do the evaluations, which stressed teaching strategies and behaviors that research linked to student success. Danielson's job was to develop a system for training the evaluators to judge teachers' strengths and weaknesses.

Praxis III was slow to catch on when it debuted in 1993 (today, only Arkansas and Ohio require candidates to pass

the evaluation to earn teaching licenses). But Danielson had noticed that the public school teachers she trained as evaluators liked the model as a way to improve teaching. So she urged ETS to adapt Praxis III for training and evaluating veteran teachers. ETS declined but gave Danielson permission to pursue the project on her own. She did, and in 1996 she published a manual, *Enhancing Professional Practice: A Framework for Teaching*.

Danielson breaks teaching down into four major categories (planning and preparation, classroom environment, instruction, and professional responsibilities), 22 themes (ranging from demonstrating knowledge of the subjects they're teaching to designing ways to motivate students to learn), and some 77 key skills (such as when and how to use different groupings of students and the most effective ways to give students feedback on their work). (See "Framework for Teaching" sidebar, page 6.) Danielson also created a set of scoring "rubrics" for evaluators that detail what teachers need to do (or not do) to earn "unsatisfactory," "basic," "proficient," and "distinguished" ratings in every skill category. (See "Scorecard" sidebar, page 8.)

The few comprehensive evaluation systems that seek to measure instruction and improve teaching use Danielson's system of standards and rubrics, or others like it. Among them are the Teacher Advancement Program, Connecticut's Beginning Educator Support and Training Program, the Cincinnati and Toledo, Ohio, school system evaluation models, and the National Board for Professional Teaching Standards.

Explicit Standards

These models share several key characteristics. The first is that they have explicit standards.

The Teacher Advancement Program (TAP) is a good example. Launched by the Milken Family Foundation

The Checklist: A Standard Teacher Evaluation Form

Teacher: _____

Evaluator: _____ Date: _____

	Satisfactory	Need Improvement
1. Knowledge of subject matter	_____	_____
2. Displays interest and enthusiasm	_____	_____
3. Shows concern for the student	_____	_____
4. Lesson preparation	_____	_____
5. Ability to motivate students	_____	_____
6. Ability to maintain student interest	_____	_____
7. Class participation	_____	_____
8. Classroom control	_____	_____
9. Respect for teacher	_____	_____
10. Poise and confidence of teacher	_____	_____
11. Brings class to initial task quickly	_____	_____
12. Majority of students on task	_____	_____
13. Has minimal interruptions in proceedings	_____	_____
14. Treats students with respect	_____	_____
15. Moves to confront problems	_____	_____
16. Seeks outside assistance if needed	_____	_____
17. Reinforces good behavior	_____	_____
18. Conducts herself/himself in a professional manner	_____	_____
19. Follows school policy/procedure	_____	_____
20. Record keeping accurate and punctual	_____	_____

Commendations/Recommendations: _____

_____

_____

_____

Source: Heart-of-the-Valley ITV, http://www.hovc.k12.nd.us/forms/Teacher%20Evaluation%20Form.doc.

in 1999 and now operated by the nonprofit, California-based National Institute for Excellence in Teaching, TAP has made Danielson's model the centerpiece of a comprehensive program to strengthen teaching through intensive instructional evaluations, coaching, career ladders, and performance-based compensation.[14] It's now in 180 schools with 5,000 teachers and 60,000 students in five states and the District of Columbia.

TAP has tweaked Danielson's teaching standards into three major categories—designing and planning instruction, the learning environment, and instruction—and 19 subgroups targeting things like how well lessons are choreographed, the frequency and quality of classroom questions, and ensuring that students are taught challenging skills like drawing conclusions.

Schools using TAP evaluate their teachers using a Danielson-like rubric that rates performance as "unsatisfactory," "proficient," or "exemplary." Standards and rubrics such as TAP's "create a common language about teaching" for educators, says Katie Gillespie, a fifth-grade teacher at DC Preparatory Academy, a District of Columbia charter school in its third year of using TAP. "That's crucial," says Gillespie.

Connecticut's Beginning Educator Support and Training Program (BEST), the nation's first—and, until recently, only—statewide evaluation system, draws heavily on the state's teachers in drafting standards.

The Connecticut Department of Education established BEST in 1989 to strengthen its teaching force by supplying new teachers with mentors and training and then requiring them in their second year to submit a portfolio chronicling a unit of instruction. The unit needs to involve at least five hours worth of teaching, to capture how teachers develop students' understanding of a topic over time, something "drive-by" evaluations can't and don't do.

State-trained scorers evaluate the portfolios from four perspectives—instructional design, instructional implementation, assessment of learning, and teachers' ability to analyze teaching and learning—using four standards: conditional, competent, proficient, and advanced. The state established committees of top Connecticut teachers to draft the standards, which were circulated to hundreds of teachers, administrators, and higher education faculty members for comment. (See "Scaling Up" sidebar, page 12.)

The nonprofit National Board for Professional Teaching Standards also has sponsored a large-scale system of teacher evaluations. It has conferred advanced certification in 16 subjects on some 63,000 teachers nationwide since its inception in 1987, using a two-part evaluation: Candidates submit a Connecticut-like portfolio and complete a series of half-hour online essays.

Charlotte Danielson's Framework for Teaching

Planning and Preparation
- Demonstrating Knowledge of Content and Pedagogy
- Demonstrating Knowledge of Students
- Setting Instructional Outcomes
- Demonstrating Knowledge of Resources
- Designing Coherent Instruction
- Designing Student Assessments

The Classroom Environment
- Creating an Environment of Respect and Rapport
- Establishing a Culture for Learning
- Managing Classroom Procedures
- Managing Student Behavior
- Organizing Physical Space

Instruction
- Communicating With Students
- Using Questioning and Discussion Techniques
- Engaging Students in Learning
- Using Assessment in Instruction
- Demonstrating Flexibility and Responsiveness

Professional Responsibilities
- Reflecting on Teaching
- Maintaining Accurate Records
- Communicating With Families
- Participating in a Professional Community
- Growing and Developing Professionally
- Showing Professionalism

Source: Adapted from *Enhancing Professional Practice: A Framework for Teaching, 2nd Edition* (pp. 3–4), by Charlotte Danielson, 2007, Alexandria, VA: ASCD.

Teams of teachers from around the country draft standards in each certification area, and hundreds of teachers, administrators, and state and federal officials comment before the standards are finalized. ETS manages the evaluation system under a contract with the National Board.

Multiple Measures

While traditional evaluations tend to be one-dimensional, relying exclusively on a single observation of a teacher in a classroom, the comprehensive models capture a much richer picture of a teacher's performance.

The National Board portfolios, for example, include lesson plans; instructional materials; student work; two, 20-minute videos of the candidate working with students in classrooms; teachers' written reflections on the two taped lessons; and evidence of work with parents and peers. That's on top of the six online exercises that National Board candidates take at one of 400 evaluation centers around the country to demonstrate expertise in the subjects they teach.

In total, National Board candidates spend between 200 and 400 hours demonstrating their proficiency in five areas: commitment to students' learning; knowledge of subject and of how to teach it; monitoring of student learning; ability to think systematically and strategically about instruction; and professional growth.

An advantage of portfolios is that, unlike standardized test scores, they can be used to evaluate teachers in nearly every discipline. National Board certification is open to some 95 percent of elementary and secondary teachers.

Teamwork

Another way to counter the limited, subjective nature of many conventional evaluations is to subject teachers to multiple evaluations by multiple evaluators.

In schools using TAP, teachers are evaluated at least three times a year against TAP's teaching standards by teams of "master" and "mentor" teachers that TAP trains to use the organization's evaluation rubrics (master teachers are more senior and do less teaching than mentors). Schools combine the scores from the different evaluations and evaluators into an annual performance rating.

TAP evaluators must demonstrate an ability to rate teachers at TAP's three performance levels before TAP lets them do "live" teacher evaluations. Then TAP requires schools using the program to enter every evaluation into a TAP-run online Performance Appraisal Management System that produces charts and graphs of evaluation results, which are used to compare a school's evaluation scores to TAP evaluation trends nationally. And every year TAP ships videotaped lessons to evaluators that they must score accurately using TAP's performance levels as a prerequisite for continuing as TAP evaluators.

In Connecticut, every BEST portfolio is scored using the program's standards by three state-trained teacher-evaluators who teach the same subject as the candidate. Failing portfolios are rescored by a fourth evaluator. As in the TAP program, scorers must complete nearly a week's worth of training and demonstrate an ability to score portfolios accurately before participating in the program.

Not surprisingly, using evaluators with backgrounds in candidates' subject and grade levels, as TAP and BEST do, strengthens the quality of evaluations. "Good instruction doesn't look the same in chemistry as in elementary reading," says Mike Gass, executive director of secondary education in Eagle County, Colo., where the district's 15 schools use TAP.

Under traditional evaluations—done as they are by principals or assistant principals—it's rarely possible to use evaluators with backgrounds in the candidate's teaching area, especially at the middle- and high-school levels, where teachers typically teach only one subject. Many evaluations, as a result, focus on how teachers teach, at the expense of what they teach. Evaluators, writes Michigan State's Kennedy, "are rarely asked to evaluate the accuracy, importance, coherence, or relevance of the content that is actually taught or the clarity with which it is taught."[15]

Subject-area and grade-level specialists, scoring rubrics, evaluator training, and recertification requirements like TAP's increase the "inter-rater reliability" of evaluations. They produce ratings that are more consistent from evaluator to evaluator and that teachers are more likely to trust.

Like TAP and Connecticut's BEST program, the National Board expands teacher evaluation roles and responsibilities beyond school principals. The board hires hundreds of veteran teachers with subject expertise, trains

them, tests them to ensure they use the board's rubrics correctly, and deploys them to scoring centers around the country, where they evaluate portfolios and assessment responses from teachers who are known only by the barcode that the board's given them. Twenty-five percent of the 18,000 portfolios and 100 percent of the 108,000 online exercises submitted to the board annually are each scored by two evaluators. In all, at least 12 people evaluate parts of each National Board application.

Some school systems, usually in collaboration with local teachers unions, have evaluation systems that rely heavily on experienced teachers to conduct "peer reviews"—in part as a way to expand the pool of evaluators.

The first and best-known such program is that in Toledo, Ohio, a heavily unionized city of 28,500 students and 2,300 teachers on the western edge of Lake Erie where the auto industry is the largest employer. Back in 1981, the president of the Toledo Federation of Teachers, Dal Lawrence, a maverick in the union movement who'd earned a master's degree in history and worked in sales before getting into teaching, struck a deal with Toledo school officials that made a districtwide team of veteran teachers responsible for evaluating every new Toledo teacher and underperforming veterans.

Under the Toledo Peer Assistance and Review program, about a dozen "consulting teachers" on leave from their

SCORECARD
Danielson Rubric for Rating Teachers' Ability to Engage Students in Learning

	Level of Performance			
Element	**Unsatisfactory**	**Basic**	**Proficient**	**Distinguished**
Activities and assignments	Activities and assignments are inappropriate for students' age or background. Students are not mentally engaged in them.	Activities and assignments are appropriate to some students and engage them mentally, but others are not engaged.	Most activities and assignments are appropriate to students, and almost all students are cognitively engaged in exploring content.	All students are cognitively engaged in the activities and assignments in their exploration of content. Students initiate or adapt activities and projects to enhance their understanding.
Grouping of students	Instructional groups are inappropriate to the students or to the instructional outcomes.	Instructional groups are only partially appropriate to the students or only moderately successful in advancing the instructional outcomes of the lesson.	Instructional groups are productive and fully appropriate to the students or to the instructional purposes of the lesson.	Instructional groups are productive and fully appropriate to the students or to the instructional purposes of the lesson. Students take the initiative to influence the formation or adjustment of instructional groups.
Instructional materials and resources	Instructional materials and resources are unsuitable to the instructional purposes or do not engage students mentally.	Instructional materials and resources are only partially suitable to the instructional purposes, or students are only partially mentally engaged with them.	Instructional materials and resources are suitable to the instructional purposes and engage students mentally.	Instructional materials and resources are suitable to the instructional purposes and engage students mentally. Students initiate the choice, adaptation, or creation of materials to enhance their learning.
Structure and pacing	The lesson has no clearly defined structure, or the pace of the lesson is too slow or rushed, or both.	The lesson has a recognizable structure, although it is not uniformly maintained throughout the lesson. Pacing of the lesson is inconsistent.	The lesson has a clearly defined structure around which the activities are organized. Pacing of the lesson is generally appropriate.	The lesson's structure is highly coherent, allowing for reflection and closure. Pacing of the lesson is appropriate for all students.

Source: From *Enhancing Professional Practice: A Framework for Teaching, 2nd Edition* (p. 85), by Charlotte Danielson, 2007, Alexandria, VA: ASCD.

classrooms for three years mentor and evaluate Toledo's first-year teachers through frequent informal classroom observations and as many as six, usually unannounced, evaluations per semester. Where possible, consulting teachers and the teachers they work with teach the same subjects, but not in the same schools. The evaluations focus on teachers' subject knowledge, professionalism, classroom management, and teaching skill.

At the end of each semester, the consulting teachers make recommendations on the dozen or so teachers under their supervision to a Board of Review comprised of five union officials and four school system administrators. The panel then votes on each teacher's status, with rulings requiring a six-vote majority. So to make it to their second year, Toledo teachers have to survive two rounds of review board voting.

Principals play adjunct roles in the evaluations, supplying consulting teachers with information on teachers' attendance and professional comportment, but leaving the classroom evaluations to the consulting teachers.

But both principals and the teacher-union committees that exist in every Toledo school can refer underperforming veteran teachers to the evaluation program, and the city's consulting teachers typically work with one or two such teachers a year. Principals or other building administrators handle the evaluations of other veteran teachers.

The Cincinnati school system and the Cincinnati Federation of Teachers have layered a Toledo-like peer assistance and review program on top of an evaluation system that uses Danielson's framework and rubrics. Principals evaluate new hires and veterans when they are up for tenure and every five years after they've won tenure. Principals conduct at least four observations during an evaluation year, and teachers who don't make the grade are frequently tracked into peer review—where a team of top teachers work with them and, if they don't improve, recommend them for dismissal. About 10 of the city's 700 teachers are placed in the program annually.

Places to Grow

Unlike traditional teacher evaluations, the systems in Toledo, Cincinnati, and Connecticut are part of programs to improve teacher performance, not merely weed out bad apples. They are drive-in rather than drive-by evaluations. At a time when research is increasingly pointing to working conditions as being more important than higher pay in keeping good teachers in the classroom, the teachers in the comprehensive evaluations programs say that the combination of extensive evaluations and coaching that they receive helps make their working conditions more professional, and thus more attractive.

At DC Preparatory Academy, which serves 275 middle-school students in northeastern Washington, D.C., using evaluations to strengthen teaching is part of the fabric of the school. The school opened in 2003 and brought on TAP in 2005. And in the TAP model, a key role of evaluations by master and mentor teachers is identifying the teachers' weaknesses that mentors will work on with teachers during the six weeks between evaluations.

"I felt I was a really good teacher before I got here," says Gillespie, in her second year at DC Prep after spending four years teaching in nearby Fairfax County, Va. "I got really high marks on my evaluations [in Fairfax]. But holy moly, I've learned under TAP that I've got a lot of places to grow." Some studies have suggested that teachers' performance plateaus after several years in the classroom. But few teachers in public education get the sort of sophisticated coaching that Gillespie receives under TAP; if more did, perhaps studies would reveal that their performance continued to improve. (See "Doing It Right" sidebar, page 17.)

"It makes a difference when people are constantly there to help you," adds Gillespie's colleague, seventh-grade English teacher Geoff Pecover. "The expectations are high. My principal last year in DCPS [the District of Columbia Public Schools, where Pecover taught for three years] showed up to evaluate my class with the evaluation form already filled out and the post-conference was a waste time. You didn't feel like you were learning anything."

To further strengthen the relationship between evaluation and instruction, TAP requires schools to have weekly, hourlong "cluster" meetings where master/mentor teachers work with teams of teachers of a particular subject or grade level.

Cost Factors—Time and Money

Not surprisingly, comprehensive classroom evaluation systems are more time-consuming and more expensive than once-a-year principal evaluations or evaluations based only on student test scores.

In schools with complex models like TAP's, the administrative challenges of training and retraining evaluators, conducting classroom visits, and tying the evaluation system to teacher professional-development activities are daunting. "We didn't realize how demanding it was," says Natalie Butler, DC Prep's principal. "You just have to make the investment."

The only way TAP, the National Board, Toledo, and other programs are able to provide multiple evaluations by multiple evaluators is by using such strategies as peer review and remote scoring of portfolios.

A few schools, including Williamsburg Collegiate, a Brooklyn, N.Y., charter school, use a two-leader model that allows one school director to evaluate and coach teachers regularly as the school's instructional leader, while a second director manages the school's non-academic operations. To compliment the instructional leader's evaluations, Collegiate brings in outside observers to conduct a daylong visit of the school once a year. They observe every teacher.

TAP and other comprehensive evaluation models also are a lot more demanding on teachers under evaluation. The upward of 400 hours some candidates for National Board certification spend in that process suggests as much, and the demands are even greater on teachers facing multiple evaluations and follow-up work under programs like TAP. "The typical teacher evaluation process puts teachers in a passive role," says Catherine Fiske Natale, a Connecticut official with the state's BEST program. "This is different." But it is not unprecedented, at least by international standards. Researchers Shujie Liu of the University of Southern Mississippi and Charles Teddlie of Louisiana State University report in a study of Chinese teacher evaluation practices that Chinese teachers are expected to observe the classes of other teachers as many as 15 times a semester and write a 1,500-word essay every semester on some aspect of their teaching experience.[16]

Toledo is spending about $500,000 out of a budget of $344 million on the peer review program this year. That's $5,000 per teacher for the 100 teachers in the program. But that figure includes the year's worth of mentoring that teachers under peer review receive. Says Lawrence of the Toledo Federation of Teachers: "It's an investment, not a cost."

TAP costs anywhere from $250 per student to $700 per student, or up to 6 percent of per-pupil expenditures. That works out to between $6,250 and $14,900 per teacher. But an average of 40 percent of that money covers performance bonuses that are built into the TAP program. (Compensation for teachers in TAP schools is based partly on their evaluation scores.) As with Toledo's program, there's intensive, yearlong mentoring for teachers built into the program's budget as well. And TAP's model includes pre- and post-evaluation conferences. TAP is less expensive where collective-bargaining contracts allow teachers more time to attend TAP meetings without having to pay them extra.

Connecticut spends about $3.7 million a year on its BEST program, or just over $2,000 per teacher. About 40 percent of that money ($800 per teacher) is spent evaluating teachers' portfolios, including training and paying stipends ($100 a day) to the 500 veteran teachers the state has score the portfolios. The other 60 percent is spent on training and supporting the 1,800 or so new teachers who go through the BEST program each year, and on central administration.

The National Board for Professional Teaching Standards is the largest and most costly of the comprehensive evaluation systems. It is a vast operation with complex shipping and security systems, a nationwide network of 400 testing centers, some 100 full-time employees and 2,500 summer-time evaluators. The organization had $42 million in annual revenues in 2006, with much of the money coming from a $2,500 fee that the board charges applicants for national certification. Since its inception two decades ago, the board has spent some $477 million to develop and operate its evaluation system, including $152 million in federal funding.

At $1,000 per teacher, it would cost $3 billion a year to evaluate the nation's 3 million teachers using a Connecticut- or National Board-like portfolio or TAP's multiple evaluations-multiple evaluators model. By way of contrast, public education's price tag has surpassed $500 billion a year, including some $14 billion (about $240 per

student) for teachers to take "professional development" courses and workshops that teachers themselves say don't improve their teaching in many instances.[17]

Yet many school systems have been reluctant to use these resources on comprehensive evaluation systems such as TAP's. "It is really difficult to get them to use Title II monies," says Kristan Van Hook, TAP's senior vice president for public policy and development, referring to the section of NCLB that funnels some $3 billion in teacher-improvement grants to the nation's school systems. "They are very reluctant to change how they spend that money. It's tied up in things like salaries for reading tutors and class-size reduction."

As a result, nearly all of TAP's expansion is being subsidized by federal grants under a U.S. Department of Education-administered Teacher Incentive Fund, a $99-million program created by Congress in 2006 to promote performance pay in public education. The federal largesse is supporting 100 of the 180 schools using TAP this year, and it will help another 50 schools adopt TAP in the coming years.

Evaluations of Evaluations

At a recent Education Sector-sponsored event on teacher evaluation at the National Press Club in Washington, D.C., Chris Cerf, deputy chancellor of the New York City Department of Education, questioned the wisdom of focusing teacher evaluations on teachers' classroom practices, asking, "Why would you look at a proxy for outcomes when you could look at the outcomes themselves?" Why use models like Toledo's and BEST, in other words, when they don't base their judgments of teacher quality on student achievement—particularly when they're more expensive?

One answer is that standardized tests, with their many limitations, provide only a partial picture of what students know and are able to do. They aren't great measures of student achievement. As a result, it's important to evaluate teachers' actual instruction—the way they work with their students in their classrooms, from their teaching techniques to the types of homework they assign.

Secondly, evidence is emerging from some of the major comprehensive evaluation models that teachers' ratings under comprehensive classroom evaluations align with their students' test scores. So, to the extent that test scores do reflect how much students are learning, high ratings under comprehensive evaluations seem to be pretty good indicators of student achievement.

A 2007 study of Connecticut's BEST program found a link between teacher ratings and test scores. Mark Wilson of the University of California-Berkeley and three co-authors, including Pecheone, compare the portfolio scores of Hartford and New Haven teachers to their students' scores on Connecticut's reading tests. They found that students of teachers earning top portfolio scores gain the equivalent of three more months of learning over the course of a school year than students of teachers who earned low scores on their BEST portfolios.[18]

Researchers Anthony Milanowski, Steven Kimball, and Brad White of the University of Wisconsin-Madison reported in 2004 that teachers with higher ratings produced greater gains in student test scores under evaluation systems using Danielson-like rubrics in Cincinnati and Las Vegas and at a Los Angeles charter school.[19]

The Classroom Assessment Scoring System, or CLASS, a new rubric-based evaluation model created by Robert Pianta of the Curry School of Education at the University of Virginia, produced comparable results in a study that Pianta and his colleagues conducted of 2,000 first-, third-, and fifth-grade classrooms to test the validity of the CLASS model.[20]

Studies of the National Board's evaluations have produced mixed results. William Sanders, a leading practitioner of value-added measures of teacher performance, concluded in a 2005 study of National Board-certified teachers—commissioned by the National Board—that his research does not "support the conclusion that, in general, students of National Board-certified teachers receive better quality teaching than students of other teachers."[21]

But two studies of National Board teachers in North Carolina suggest that the board's evaluation system generally bestows the organization's imprimatur on the right teachers. A study led by Dan Goldhaber of the University of Washington and another led by Helen Ladd (and colleagues) of Duke found that the National Board model is, in Goldhaber's words, "a good sorter."[22] A recent study by Douglas Harris of the University of Wisconsin-Madison and Tim Sass of Florida State University found

that students of National Board-certified teachers in Florida outperformed those of non-board-certified teachers in some grades and subjects but not others, and more so on one Florida test than another.[23]

The relatively small number of studies linking teacher ratings under comprehensive evaluation systems to student test scores, together with the National Board's mixed report card, has led some education experts to question the value of investing in such systems. As Harris and Sass say about National Board evaluations, "In addition to the potential benefits, it is important to consider the substantial costs that go into the certification—teacher time, NBPTS administration and direct financial incentives."[24] Education experts are also skeptical of the small percentages of teachers who get low marks under some of the

Scaling Up: Connecticut Beginning Educator Support and Training Program

Connecticut has been in the forefront of teacher evaluation since 1989, when it created the nation's first statewide teacher evaluation system to help raise the quality of teachers being licensed in the state.

The Connecticut Beginning Educator Support and Training Program (BEST) combines two years of mentoring and training for every new Connecticut teacher with an evaluation of the teacher's performance against statewide instructional standards. Teachers must earn a satisfactory rating on the evaluation in order to become fully licensed.

The evaluations are based on portfolios that second-year teachers assemble to chronicle their instruction and their students' learning in a unit of instruction five to eight hours long. The portfolios include a description of the demographics and academic background of the class, unit goals, daily logs of activities, student work with teacher feedback, and a reflective analysis. Teachers must also submit a videotape of at least 20 minutes of instruction to supplement the written documents.

Each portfolio is scored by three state-trained evaluators, and failing portfolios are rescored by a fourth evaluator. Scorers must go through four days of training and demonstrate an ability to score portfolios accurately before participating in the program. They then spend about two weeks during the summer scoring portfolios, earning stipends of $100 a day.

The scorers evaluate the portfolios from four perspectives: instructional design; instructional implementation; assessment of learning; and teachers' ability to analyze teaching and learning.

The scorers, who are experienced classroom teachers from the same discipline area as the beginning teachers, rate teachers' abilities in each of the four areas on a scale of 1-to-4; 1 represents "conditional" performance, 2 is "competent," 3 "proficient," and 4 "advanced."

Scorers then assign an overall rating, again on a 1-to-4 scale. A score of 2 or above is required for full licensure. Those who receive a score of 1 on their portfolios are able to take part in the BEST program for a third year and receive additional mentoring and submit another portfolio. Those who don't earn a passing score the second time through the process are no longer eligible to teach in Connecticut public schools.

Some 1,800 teachers a year take part in the BEST program, at an annual cost to the state of $3.65 million, or just over $2,000 per teacher.

Eighty-eight percent of the portfolios received by Connecticut officials in 2003 and 2004 received passing scores. And because teachers are able to spend an additional year in the program and submit a second portfolio, only 44 out of 3,544 teachers failed the evaluation during those years.

Researchers say it's impossible to distinguish the impact of the BEST evaluations on teacher quality from the mentoring and other steps Connecticut has taken to bolster teaching in its public schools. But state officials nonetheless have been pleased with the program. Scores on the portfolios are high, says Catherine Fiske Natale, director of educator support and assessment for the Connecticut State Department of Education, because teachers understand the state's standards and have incorporated them in their instruction, and because the mentor teachers who support new teachers as part of the program are skilled. "We've seen performance levels [of teachers who are mentored] shoot up," says Natale. "We look at that as a sign of success."

Teachers in the program are also enthusiastic about it.

More than 90 percent of beginning teachers tell state education officials that the BEST program improves their teaching. The teachers who participate in BEST as mentors and evaluators also give the program equally high marks. In surveys, 80 percent of mentors and 90 percent of scorers say their roles in BEST have made them better teachers. Becky Wentworth, a fifth-grade teacher at Windermere School in Ellington, says that evaluating the Connecticut portfolios enables teachers to re-examine their practices and work toward improving them. "It helps because I look at my teaching differently," she says. "I'm more thoughtful about my own work."

There's some evidence that the program also has helped in other ways. A 2002 study by Michigan State University professor Peter Youngs found that the BEST program and higher teacher salaries combined to keep teacher attrition in the Bristol and New Britain school districts low. Some 87 percent of teachers in Bristol and 91 percent of those in New Britain stay through their first three years, compared to as few as 70 percent in some urban districts nationally.*

New Mexico and Wisconsin recently have introduced portfolio-based evaluations of new teachers similar to Connecticut's.

*Peter Youngs, "State and District Policy Related to Mentoring and New Teacher Induction in Connecticut," (paper prepared for the National Commission on Teaching and America's Future, December 2002).

comprehensive evaluation systems. Some 37 percent of National Board candidates are successful the first time they apply for national certification, and many reapply, so about two-thirds of the board's applicants eventually earn certification.

In Toledo, about 10 percent of new teachers fail peer review and leave the school system, not a small percentage. The number of veterans dismissed through the program is much smaller. In 27 years close to 100 veteran teachers out of the city's 2,300-teacher workforce have been fired through peer review—though some failing teachers do quit when it's clear they're headed for dismissal.[25]

A meeting of the program's review board in downtown Toledo in the fall of 2007 suggests that peer review turns up plenty of problem teachers. With the review board's four administrators and five union representatives seated around a horse shoe-shaped table to hear progress reports on some two dozen teachers who had started the year badly under peer review, the program's consulting teachers presented the panel with a litany of troubled classrooms and bad instruction in history, special education, woodworking, kindergarten, and math. There are a total of 90 teachers under peer review this year.

Three of the teachers discussed during the meeting resigned in the following weeks. The board terminated a fourth at a subsequent meeting in January of 2008 and gave half a dozen others unsatisfactory ratings for their first semester at the same meeting. By year's end, about half of the teachers on the docket during the fall would be out of the Toledo school system, predicted Carol Thomas, the Toledo Public Schools' director of human resources and an administrative representative on the review board.

Thomas says the program "is much more rigorous" than Toledo's evaluations of veteran teachers, which are done by building administrators and are "more of a judgment call." The presence of peer preview "has produced a conversation within the union about teaching quality," says Lawrence of the Toledo Federation of Teachers, adding that under Toledo's traditional evaluations, "we never fired anyone."

In Cincinnati, "more people have been recommended for non-renewal through peer review than administrator evaluation," says Tim Kraus, the president of the Cincinnati Federation of Teachers, noting that the coaching veteran teachers receive under the program reduces the number of low ratings by design.

Officials in Connecticut make the same argument. Ten percent of the state's teachers failed their BEST portfolio evaluations in 2006–07 on their first try. Of those who passed, 50 percent were rated "competent," 30 percent "proficient," and 10 percent "advanced." But the failure rate eventually drops to about 2 percent under BEST because many teachers avail themselves of an opportunity to get another year's worth of mentoring and then go through a portfolio evaluation a second time.

The National Institute for Excellence in Teaching reports that TAP evaluators gave ratings below "proficient" to about 20 percent of the nearly 490 South Carolina teachers evaluated under TAP in 2006–07, rendering the teachers ineligible for a portion of TAP's performance pay. Of the 227 teachers who left TAP schools nationally during or after the 2006–07 school year, 27 percent left involuntarily. Not all those teachers departed because of bad evaluations, but certainly the percentage that did is substantially higher than the infinitesimal percentages of departures that result under traditional evaluation models.

Sending a Message

Comprehensive evaluations—with standards and scoring rubrics and multiple classroom observations by multiple evaluators and a role for student work and teacher reflections—are valuable regardless of the degree to which they predict student achievement, and regardless of whether they're used to weed out a few bad teachers or a lot of them. They contribute much more to the improvement of teaching than today's drive-by evaluations or test scores alone. And they contribute to a much more professional atmosphere in schools.

As a result, they make public school teaching more attractive to the sort of talent that the occupation has struggled to recruit and retain. Capable people want to work in environments where they sense they matter and using evaluation systems as engines of professional improvement signals that teaching is such an enterprise. Comprehensive evaluation systems send a message that teachers are professionals doing important work.

There's always going to be some degree of subjectivity in evaluation of work as complex as teaching. But TAP

and other comprehensive systems using standards and rubrics and multiple evaluators are sufficiently objective to be credible in teachers' eyes.

"It's fair, and the consulting teachers are nothing but helpful," says Mike Blackwood, a geologist turned chemistry teacher at Toledo's Libbey High School. Says Blackwood, who went through Toledo's peer-review program as a first-year teacher in 2006–07: "They gave me ideas about everything from how to use manipulatives to classroom management."

In Toledo Federation of Teachers surveys, 90 percent of Toledo teachers support the city's peer-review system, even though the program violates the traditional union principle that only management should evaluate labor. The Cincinnati Federation of Teachers also reports "overwhelmingly positive" support for its evaluation system among its members, says Lesley-Ann Smillie, the organization's professional issues representative.

"It's definitely fair," says Rosemary Penna, a science teacher at Silver Spring International Middle School in Montgomery County, Md., of the National Board evaluation systems. "The rubrics make the process transparent. It's subject-specific. There are lots of evaluators." Penna earned the board's science certification in December 2007.

Though Gillespie of DC Prep hasn't earned high marks under the TAP evaluation system, she likes it. "It's not subjective," she says. "It's, 'You did this, you didn't do that, and here's the result.' I trust it."

TAP, the National Board and other systems that rely on teachers to conduct evaluations also create a more professional environment for teacher-evaluators.

"Examining what other teachers do and comparing it to standards causes you to be much more reflective about your own teaching," noted JoEllen Belter, a reading specialist at North Canaan Elementary School in North Canaan, Conn., during a break from scoring BEST portfolios at an East Hartford high school last fall. "In every portfolio where you identify an area of weakness, it causes you to reflect, 'What would I have done?'"

Teacher-evaluators also enjoy a step up in status and pay under BEST, the Toledo plan, and other systems. TAP master and mentor teachers become part of their schools' leadership teams.

The importance of a professional working environment to many teachers is reflected in a 2007 national survey of teachers by the nonprofits Public Agenda and the National Comprehensive Center for Teacher Quality (NCCTQ). The organizations found that if given a choice between two otherwise identical schools, 76 percent of secondary teachers and 81 percent of elementary teachers would rather be at a school where administrators supported teachers strongly than at a school that paid significantly higher salaries.[26]

Ken Futernick, the director of the Center for Teacher Quality at the California State University-Sacramento, found the same sentiment in a 2007 survey of 2,000 current and former California public school teachers. They, too, stressed working conditions over compensation in deciding whether to leave the teaching profession.[27]

Performance Pay

But 70 percent of the teachers in the Public Agenda/NCCTQ survey also saw public education's failure under the single salary schedule to reward teachers "for superior effort and performance" as a "drawback to [public school] teaching." Younger teachers, in particular, want to work in an environment that rewards performance.

Yet teachers don't trust either principals by themselves or test scores to reward performance fairly. In a 2007 report on teacher attitudes about compensation reform, Goldhaber and colleagues at the nonprofit Center for Reinventing Public Education note that only 3 percent of teachers in a national poll conducted several years ago were willing to use student test scores as a factor in determining teacher salaries.[28]

But many teachers are willing to be part of performance-pay systems when ratings are based substantially on comprehensive evaluations of classroom performance.

In TAP, where compensation is based partly on evaluation scores, 40 percent to 50 percent of teachers' ratings are based on their classroom evaluations, and the rest

are divided between their students' achievement gains and schoolwide results. TAP only uses value-added calculations of test scores, and its ratings of teachers of non-tested subjects are based on a combination of classroom evaluations and schoolwide scores.

In TAP surveys, only a third of teachers in TAP schools say performance pay is a negative influence on their schools, compared to twice that percentage in other national surveys.[29] "Test scores alone aren't the answer," says Lowell Milken, TAP's creator and chairman of the Milken Family Foundation. "Multiple measures [for determining performance pay] are important because the tests aren't perfect." Milken and others in the TAP system also say that TAP's linking of its evaluations to classroom coaching by master and mentor teachers is equally important to winning teachers' support of TAP's performance-pay plan.

In Toledo, about 5 percent of the city's 2,300 teachers participate in a voluntary performance-pay plan that's an offshoot of the city's peer review program. They earn 15 percent salary increases by passing six evaluations by three consulting teachers and by giving up seniority to work in hard-to-staff schools. "We wouldn't have it without PAR [the peer assistance and review program]," says Lawrence of the Toledo Federation of Teachers.

The Denver Professional Compensation System, known as ProComp, is one of the most closely watched teacher performance-pay experiments in the country. But it, too, rewards teachers only partly on the basis of their students' test scores. Just as important are exemplary evaluations by school principals (using scoring rubrics) and successfully introducing new teaching strategies.[30] Another part of the plan rewards teaching hard-to-staff subjects or in hard-to-staff schools.

In sharp contrast, Florida in 2002 launched a statewide performance-pay plan that handed out bonuses of about $2,500 to 10 percent of the state's teachers, based solely on their students' standardized test scores. Teachers with top-scoring students got the money, regardless of the students' academic backgrounds, while teachers who taught untested subjects were excluded from the pay plan.

The Florida E-Comp program created a furor within the state's teaching ranks, and in 2006 the Florida Legislature, running for cover from the enraged educators and their powerful union, expanded the program's performance pay quota from only 10 percent to 25 percent of the state's teachers with the highest student test score gains.

The strategy, however, didn't work, and in 2007 the state's lawmakers revamped the program a second time, making it voluntary for school systems, permitting schoolwide as well as individual awards, and requiring that 40 percent of the calculations for the awards be based on teachers' classroom evaluations. But they also de-funded the $148-million program for a year to help patch a hole in the state's budget, which further diminished the program's credibility in the eyes of the state's teachers.

In the 1980s, in the wake of calls for performance pay by the authors of "A Nation at Risk" and other reform manifestos, Florida, other states, and scores of school systems committed the same mistakes that Florida has made more recently: providing token bonuses to arbitrary numbers of teachers on the basis of subjective standards (for the most part, principals picked the award winners) under funding that rose and fell with state fiscal tides. Virtually none of the programs survived into the 1990s.

Teachers Unions

Ultimately, the expansion of comprehensive evaluation systems depends on teachers unions' willingness to back them, because the unions exert tremendous influence over teacher policies at every level of education policymaking, even in states without collective-bargaining laws.

But the unions have not, in the main, sought to improve the unproductive ways that teachers are evaluated in most school systems today.

Back in 1985, Albert Shanker, the powerful president of the American Federation of Teachers (AFT), the nation's second-largest teachers union, made a compelling case for union support of rigorous evaluations. "We don't have the right to be called professionals—and we will never convince the public that we are," he told a union convention in Niagara Falls, "unless we are prepared honestly to decide what constitutes competence in our profession and what constitutes incompetence and apply those definitions to ourselves and our colleagues."[31]

But for many local, state, and national union leaders the impulse to protect the jobs of their members has outweighed Shanker's broader view of how to improve the status of public school teachers. They have not pressed for more rigorous evaluation systems for fear that such systems may result in more teachers being dismissed for poor performance, and strengthen the case for performance-based pay at the expense of the single salary schedule.

"The public education culture is so deeply rooted in the industrial model [of labor-management relations] that union people and administrators find it extremely difficult to be proactive," says Lawrence of the Toledo Federation of Teachers. "The natural response is to be on the defensive. Defend. Defend. Defend. What I hear most often from union leaders is that it's management's job to evaluate. It's sad."

"Too many people think checklists are just fine," adds Susan Carmon, associate director of the teacher quality department at the National Education Association (NEA), which represents about two-thirds of the nation's teachers (the AFT represents about one-fifth). "We do not have effective systems of teacher evaluation. They have the potential to be volatile in labor-management relations, so people are reluctant to jump with both feet" to strengthen them.

In a classic example of the conflicts that rigorous teacher evaluation create for teacher unions, the United Teachers of Dade (UTD), the union representing teachers in the Miami area, back in the 1980s endorsed a proposal to replace the school system's cursory teacher evaluation checklists with a more comprehensive system designed to give teachers feedback on their performance and to weed out weak performers more effectively. In response, the UTD's rival, the local affiliate of the NEA, famously took out a full-page ad in the *Miami Herald* charging that the UTD was undermining teacher job security.[32]

The AFT points with enthusiasm to the comprehensive evaluation systems that some of its local affiliates have implemented. In St. Francis, Minn., a district of 6,100 students 30 miles from Minneapolis, the union and school system have introduced a model calling for a "professional review team" of two teachers and an administrator to evaluate every St. Francis teacher four times a school year, using a rubric-based system with pre- and post-evaluation conferences. Three years' worth of satisfactory evaluations earns teachers salary increases funded in part by Quality Compensation for Teachers, or Q-Comp, a teacher-improvement initiative launched by Minnesota lawmakers in 2005.

The AFT also is urging its affiliates to initiate Toledo-like peer review programs. Executive Vice President Antonia Cortese praised peer review in a speech to the AFT membership in 2007, and the organization has launched a project to incubate peer review programs in a number of school systems, hiring Lawrence to promote the effort.

But the two major national unions can't dictate the policies of their locals, and there have been few takers for peer review: 27 years after Toledo originated the model, only 50 or 60 of the nation's 14,000 school systems are using it.

Some principals see peer review as a union power grab, says Joan Devlin, the AFT official heading the new initiative, while many union leaders think it violates their obligation to represent their members' interests. Such was the unease that the Toledo plan created on both sides of the bargaining table that shortly after the plan's launch the Toledo Federation of Teachers had the Ohio Legislature write a clause into the state's teacher-bargaining law protecting the bargaining rights of Toledo teachers evaluating their peers, because teacher evaluations had always been seen as management work.

Both the NEA and the AFT are strongly against using student test scores to evaluate individual teachers. The NEA's powerful delegate assembly passed a resolution in 2007 declaring that "the use of student achievement measures … to determine the 'competency, quality, or effectiveness' of any teacher is "inappropriate."[33] The California Teachers Association, an NEA affiliate, in 2006 won a legislative prohibition against the use in teacher evaluations of a new statewide system for tracking California student test scores.

And the NEA and the vast majority of its state and local affiliates oppose performance pay based on either classroom evaluations or student test scores. "The Association … believes that … compensation based on an evaluation of an education employee's performance" is "inappropriate," says the NEA in a

resolution, adding that "Any additional compensation beyond a single salary schedule must not be based on education employee evaluation, student performance, or attendance."[34]

The union has declared publicly its unhappiness with its Denver affiliate's sponsorship of the city's ProComp performance-pay system. And the NEA's leadership attacked Rep. George Miller, the liberal Democratic chairman of the House Education and Labor Committee (and a natural NEA ally), last fall over performance-pay provisions that Miller had included in a draft of legislation to reauthorize NCLB. The AFT also clashed with Miller over the plan's use of test scores to rate teachers.

The AFT and some of its affiliates, including in St. Francis, have been open to performance pay plans that don't target individual teacher's test scores. Last fall, the AFT's largest affiliate, the United Federation of Teachers (UFT), which represents New York City's teachers, agreed to a two-year pilot program that awards performance bonuses to schools rather than individuals, a model that the NEA is more tolerant of.

TAP encourages schools to win the support of 75 percent of its teachers before joining the TAP network, as a way of heading off opposition to its hybrid performance-pay system that evaluates teachers on the basis of both classroom evaluations and student test scores.

Unions deny their members an opportunity to grow professionally when they oppose comprehensive evaluation systems like TAP's. But regardless of the evaluation system, teachers aren't going to buy into a performance-pay system that pegs a substantial percentage of their compensation to their performance evaluations. Unlike on Wall Street, where large sums of performance pay often are stacked on top of already-generous base salaries, teachers, who earn an average of about $50,000 a year in the United States, want the majority of their pay in the form of a fixed annual income.

That's one reason why the members of the Cincinnati Federation of Teachers in 2002 rejected by a vote of 1,892 to 73 a performance-pay plan based on the city's Danielson-inspired classroom evaluation system. Teachers with bad evaluations risked moving down the city's pay scale.

Doing It Right: Teacher Advancement Program

In schools using the Teacher Advancement Program, teachers are evaluated at least three times a year against TAP's teaching standards by teams of "master" and "mentor" teachers that TAP trains to use the organization's evaluation rubrics.

Last fall at DC Preparatory Academy, a charter middle school in the District of Columbia, master teacher, Mary Kate Hughes, and mentor teacher Cassie Meltzer met with fifth-grade teacher Katie Gillespie in her empty classroom to discuss a reading lesson Gillespie would teach later in the day on the value to readers of making predictions about what's likely to happen next in the stories they are reading. The meeting was the first step in Gillespie's first formal evaluation of the school year.

Working off of a standard form that she had completed, Gillespie walked her colleagues through the lesson she planned to teach, detailing what she wanted to accomplish and how she planned to do it.

Hughes and Meltzer peppered her with questions. Was she making sure she explained why predictions are an effective reading strategy? What's the difference between a good prediction and a mediocre one? How would she make the distinction clear to her students? Meltzer suggested that she refer to meteorologists on the local television news to make the point that predictions are sometimes right, sometimes wrong.

An hour later, Meltzer and Hughes and a third TAP-trained evaluator, administrator Katie Severn, sat in the back of Gillespie's class as she moved through a "mini-lesson," a "read-aloud," silent reading, and work by students working in teams—all designed to teach students the value of readers making predictions. Meltzer, Hughes, and Severn took volumes of notes on everything that transpired in the room.

Afterward, Severn and Meltzer met in the DC Prep cafeteria to debrief. They tallied the strengths and weaknesses of Gillespie's lesson, landing on several things for her to work on: ensuring that classroom tasks are sequenced properly (Gillespie talked about the importance of making predictions and then had her students do silent reading without first giving them examples of different types of predictions); modeling more clearly for students how they should go about a task; and holding students accountable when they break class rules.

Meltzer would stress these skills in the regularly scheduled coaching she'd do with Gillespie until Gillespie's next evaluation. Under TAP, every teacher has a mentor like Meltzer. They're a constant presence in classes—taking notes, teaching model lessons, recommending reading materials, organizing observations of colleagues' classes.

By the next day, Meltzer and Severn had written up their comments and discussed them with Gillespie, who shared her self-evaluation of the predictions lesson. The meeting ended with praise for Gillespie's strengths and a plan for improvement.

New Incentives

It's hard to believe that an industry that spends $400 billion annually on something as central to its success as teachers are to public education pays so little attention to the return on its investment. How can public education hope to improve teacher quality without a reliable way to measure teacher quality?

Teacher ratings based on student standardized test scores aren't, by themselves, the answer. Sure, they're cheaper, simpler evaluation tools. They seemingly measure what matters most—student achievement. And they are a hedge against the degree of subjectivity that exists in even the most comprehensive classroom evaluations.

But the partial picture they paint of student achievement, and the fact that they leave a blank canvas for the many teachers who don't teach tested subjects, argues that they not play a lead role in teacher evaluations.

To get a fuller and fairer sense of teachers' performance, evaluations should focus on teachers' instruction—the way they plan, teach, test, manage, and motivate. They need to move far beyond principal drive-bys to multiple measures, multiple evaluations, and multiple evaluators. And they should contribute to helping teachers improve their performance to a far greater degree than they do in most public schools today—both to promote a climate that attracts the best and brightest into teaching and to spend public education's vast "professional development" monies far more efficiently than most school systems do today.

Where possible, in the most defensible ways possible, student test scores should have a role in teacher evaluations. School systems should evaluate both the work that teachers do in their classrooms and the results of that work. As Joan Baratz-Snowden, a former director of educational issues at the American Federation of Teachers, says: "Anyone who dismisses student learning [in evaluations] is naïve. Anyone who defines student learning as tests scores is also naïve."

But test scores should have a minor role, accounting for under 50 percent of a teacher's evaluation. And school systems should use schoolwide scores in their evaluation calculations, rather than individual teachers' scores. That's because many teachers don't teach tested subjects, the small number of students that many teachers teach skews the results, and using schoolwide scores encourages school staffs to collaborate rather than compete.

But superficial principal drive-bys will continue to pervade public education—and teacher evaluation's potential as a lever of teacher and school improvement will continue to be squandered—if school systems and teachers unions lack incentives to do things differently.

NCLB has helped. By creating consequences for schools and school systems with students who fall below state standards, the law "is pushing principals" to take evaluations more seriously, says Flannery of the secondary school principals association. In Toledo, says Thomas, the school system's director of human resources, both principals and teacher building teams are referring more veteran teachers to peer review in the wake of NCLB "because they don't want to work with people who are pulling the whole school down."

New York City's school system, the nation's largest, recently layered on top of NCLB a system of sanctions and rewards for both schools and their principals that gives teachers and principals alike strong incentives to care about the quality of the teaching in their classrooms.

Giving schools facing such carrots and sticks greater authority over teacher hiring and firing would further incentivize them to evaluate teachers carefully.

Ultimately, the single salary schedule may be the most stubborn barrier to better teacher evaluations. As Kate Walsh, president of the National Council on Teacher Quality and member-designate of the Maryland State Board of Education, says: "If there are no consequences for rating a teacher at the top, the middle or the bottom, if everyone is getting paid the same, then why would a principal spend a lot of time doing a careful evaluation? I wouldn't bother." Many teachers unions, of course, argue that the failure of principals to take evaluations seriously requires a single salary schedule.

There's no simple solution to this Catch 22. But TAP, for one, has addressed it head-on by combining comprehensive evaluations that teachers trust with performance pay. The program's comprehensive classroom evaluations legitimize performance pay in teachers' minds, and its performance-pay component gives teachers and administrators alike a compelling reason to take evaluations seriously. Pay and evaluations become mutually reinforcing, rather than mutually exclusive.

Recommendations

There are several steps that federal, state, and local policymakers should take to strengthen teacher evaluation in ways that would help school systems to judge teachers' strengths and weaknesses more fairly and effectively and to use evaluations to improve teaching.

A Hybrid Model

Evaluate teachers on the basis of instruction and student achievement: The vast majority of school systems today evaluate teachers strictly on the basis of a single, ad hoc classroom visit by a teacher's principal. That's in part because teaching has long been seen as more art than science.

But the experience of standards-based evaluation systems suggests that there are identifiable teaching traits that raise student achievement. So classroom observations should continue to be central to teacher evaluations, but they need to become much more sophisticated.

Evaluations should be based on teaching standards accompanied by rubrics for measuring teachers' success. As Joan Baratz-Snowden, a former director of educational issues at the AFT, says: "Evaluation without standards becomes a matter of taste. There has to be a shared understanding of what quality teaching is."

Evaluators should be trained in the use of standards and rubrics, and evaluations should be based on multiple classroom observations by multiple evaluators. It's hard to overstate teachers' resentment of what they believe to be the inherent arbitrariness of single-evaluator evaluations. As so-called 360-degree evaluations become increasingly widespread in other fields, it makes sense to include surveys of students and parents in teacher evaluations. The New York City school system has begun including such surveys in its school report cards.[35]

The experiences of the leading comprehensive evaluation systems suggest that samples of student work, teachers' assignments, and other "artifacts" of teaching are valuable compliments to classroom observations and should be included in evaluations. A 2001 study by the Consortium on Chicago School Research at the University of Chicago found a strong correlation between the rigor of teachers' math and reading assignments and their students' standardized test scores.[36]

Because so many teachers don't teach tested subjects, and because most standardized tests today measure only a narrow range of mostly low-end skills, test scores should account for less than half of teachers' evaluation ratings, and they should be based on schoolwide increases in students' scores during a school year. Using averages would also encourage cooperation among teachers rather than competition.

Beyond Drive-Bys

Trained Evaluators: Not surprisingly, standards-based evaluations are effective only if evaluators know how to use them effectively. But most principals are poorly trained in teacher-evaluation techniques.

In contrast, TAP, BEST, and other comprehensive programs train their evaluators extensively—both to ensure that evaluators use standards and rubrics accurately, but also to ensure that ratings are consistent from one evaluator to the next. This "inter-rater reliability" is difficult to achieve, but important to teacher morale. Multiple classroom observations also increase teachers' trust.

District Evaluation Teams: Because principals lack the time and the training to conduct comprehensive teacher evaluations, and because many experts say that principals are reluctant to evaluate rigorously teachers they work with every day, school systems should create cadres of trained district-level evaluators of the sort that Toledo has established under its peer review program.

These evaluation teams should be comprised of assistant principals, as part of their preparation to become school leaders, and teachers, who earn the right to be evaluators as a reward for outstanding teaching (and the assignment should come with the title of master teacher and a salary increase). Both administrators and teachers on district evaluation teams should serve full-time for at least a year and preferably longer.

A trained cadre of evaluators would produce more objective and more consistent evaluations. Connecticut, for example, requires teachers scoring BEST portfolios to recuse themselves from scoring the portfolios of teachers that they know. As a national program, TAP checks the consistency of local evaluations by requiring that those who do evaluations be re-tested and re-certified as TAP evaluators every year. TAP does this by sending its schools a DVD of a taped lesson, and the local evaluators must score the teaching performance accurately.

But if principals are going to be instructional leaders, they need to have a feel for the instruction in their schools. So they should contribute to the evaluation of the teachers in their buildings, working closely with district evaluation teams.

An Out-sourcing Option: Another option would be to outsource some of a teacher's classroom "observations" using a portfolio model similar to BEST's in Connecticut. Larger school systems and smaller states could establish scoring programs like BEST's that would offer trained evaluators, objectivity, and reduced administrative costs. A new portfolio model that a consortium of California colleges and universities plan to start using in 2008–09 to evaluate some 14,000 probationary teachers annually is expected to cost about $400 per teacher.

Evaluate the Evaluations

More research is needed on the predictive validity of comprehensive teacher-evaluation models, particularly national programs such as TAP. And studies of National Board certification should be conducted in states like Massachusetts and Connecticut, states with tests that measure a wider range of skills and knowledge than do tests in most other states.

School, School Leader Incentives

School and Principal Rewards: Creating school system evaluation teams sends a strong signal to principals and teachers that evaluations—and performance—are high priorities. Districts could reinforce that message by establishing a system of performance-based rewards and sanctions for schools and school leaders.

These carrots and sticks should be based on a range of factors that include parent and student surveys and test scores to stress the importance of the quality of classroom instruction. And districts should rate schools on how they stack up against peers with similar demographics, to make the comparisons fair.[37] New York City's new performance bonus program is a good model; it includes both financial rewards and administrative sanctions.

Staffing Authority: Giving principals a greater say in selecting their teachers and in dismissing those who don't perform would further strengthen school leaders' incentives to make evaluations matter. If principals think they have a significant stake in staffing decisions—and if they're on the hook for the results—they're going to be more invested in teacher evaluations.

Evaluations and Professional Development: State lawmakers and local school boards should require that the $14 billion public schools spend annually on professional development be targeted to addressing individual teacher weaknesses identified through comprehensive teacher evaluations. Currently, much of the money is spent on college courses that often have little relevance to teachers' classrooms. Tying professional development directly to the gaps in teachers' skills would be a far more efficient way to spend those monies.

Performance Pay

Public school teachers earn an average of about $52,500, not enough to base a substantial portion of their pay on performance: People aren't going to enter or stay in a profession that puts such low pay at risk. So performance pay should constitute a relatively modest percentage of teachers' compensation: not too small to be meaningless; not too large to drive people away from the profession. Teachers with high performance ratings should also have opportunities to become mentor or master teachers and earn higher salaries for those roles. Such modifications to the single salary schedule would draw greater attention to the quality of teacher evaluations. And programs like TAP demonstrate that strong evaluations and performance can be mutually reinforcing.

Denver pilot-tested its ProComp performance-pay system in a small number of schools and commissioned a study to improve the experiment before launching it districtwide, a strategy that won the program support among teachers and the Denver community.

A Federal Role

A New Definition of "Qualified" Teachers: To help leverage change, Congress should modify NCLB to require that all public school teachers earn a designation as "highly qualified *effective* teachers." The Aspen Commission proposed such a step, but urged that student test scores play a leading role in defining teacher effectiveness. We believe that test scores should play a less significant role and that any federal definition of "highly qualified" or "effective" teachers should include criteria that allow states to innovate with comprehensive standards-based approaches to teacher evaluation. Such flexibility, enshrined in federal law, would encourage states and school systems to focus on teacher performance rather than teacher credentials and to take teacher evaluations far more seriously than they do now.

Appendix. Standards-Based Teacher Evaluation Models

New Teacher Assessments

PACT

Performance Assessment for California Teachers (PACT) was developed by a consortium of 30 California colleges and universities (and one district-run preparation program) that trains about 30 percent of the state's 20,000 new teachers every year. The candidates submit a portfolio that includes lesson plans, "artifacts," a teaching videotape, and reflections on teaching from a three-to-five-day period during their student teaching. The portfolios are evaluated by trained scorers who focus on four areas: planning, instruction, assessment, and reflection. Beginning in 2008–09, teacher-candidates from the consortium institutions must pass the assessment in order to earn a license. The program is expected to cost approximately $400 per teacher.

For more information: Raymond Pecheone, School Redesign Network, Stanford University, School of Education. pecheone@stanford.edu.

Praxis III

Praxis III measures the teaching skills of teacher-candidates. It's part of a series of teacher-licensure examinations developed by the Princeton-based Educational Testing Service that also includes Praxis I, a basic-skills test, and Praxis II, a series of tests of subject-matter and pedagogical knowledge. It measures teachers' abilities in four areas: organizing content knowledge, creating a classroom environment conducive to learning, instruction, and teacher professionalism. The in-class evaluations are conducted by ETS-trained evaluators who do pre- and post-evaluation conferences with teachers. Teachers in Arkansas and Ohio must pass Praxis III to earn teaching licenses.

For more information: Educational Testing Service. http:// www.ets.org/portal/site/ets/menuitem.

BEST

The Connecticut Department of Education created the Beginning Educator Support and Training (BEST) program in 1989 to ensure that the state's new teachers provided effective classroom instruction. Second-year teachers submit a portfolio that includes the teaching materials for a lesson, examples of student work, reflections on teaching the lesson, and a video of themselves teaching. The state trains teachers from throughout Connecticut to score the portfolios at state-run scoring centers.

For more information: Connecticut State Department of Education, Bureau of Educator Preparation, Certification, Support and Assessment. (860) 713-6543; http://www.sde.ct.gov/sde/cwp/view.asp?a=2607&Q=319186.

Standards-Based Evaluations of Practicing Teachers

CLASS

Researchers at the Center for Advanced Study of Teaching and Learning at the University of Virginia developed the Classroom Assessment Scoring System (CLASS) as a tool to evaluate teachers of students in pre-k and lower elementary grades. It includes observations of classrooms by trained evaluators who measure teachers' performance in three areas: emotional support (including classroom climate, teacher sensitivity, and student perspectives); classroom organization; and instructional support (including quality of feedback and language modeling). It is currently being used to assess and provide support for preschool teachers in Wyoming and Massachusetts, and as part of a web-based professional development system called MyTeachingPartner. The American Board for Certification of Teaching Excellence is using the system in a pilot program to recognize Distinguished Teachers.

For more information: (866) 301-8278; http://www.classobservation.com.

The Toledo Plan

The Toledo Public Schools and the Toledo Federation of Teachers since 1981 have co-sponsored a "peer assistance and review" program that uses trained public school teachers to conduct comprehensive, yearlong evaluations of teachers new to the school system and underperforming veterans.

For more information: Toledo Federation of Teachers. 419-535-3109; http://www.tft250.org/peer_ review.htm.

Appendix. Standards-Based Teacher Evaluation Models (continued)

National Board for Professional Teaching Standards

The nonprofit National Board for Professional Teaching Standards has conveyed advanced certification on some 63,000 teachers in 16 subjects since its inception in 1987. The board's evaluations are based on a portfolio of videotapes and other materials that capture teachers' classroom instruction and a series of six, subject-specific online teaching exercises that candidates complete at NBPTS centers nationally.

For more information: National Board for Professional Teaching Standards. 800-228-3224; http://www.nbpts.org.

Incentive Plans

Teacher Advancement Program

Launched by the Milken Family Foundation in 1999 and now operated by the nonprofit, California-based National Institute for Excellence in Teaching, the Teacher Advancement Program (TAP) includes rigorous, standards-based evaluations by trained master and mentor teachers as part of a schoolwide performance-based compensation system. The program is now in use in some 180 schools nationwide.

For more information: National Institute for Excellence in Teaching. 310-570-4860; http://www.talentedteachers.org.

ProComp

ProComp, an incentive pay system created by the Denver Public Schools and the Denver Classroom Teachers Association, promotes teacher evaluation and professional development through increased pay tied to high performance ratings, the meeting of student-achievement objectives, and demonstration of improved knowledge and teaching skills.

For more information: 720-423-3900; www.denverprocomp.org.

Q-Comp

The Minnesota legislature created a statewide Quality Compensation for Teachers (Q-Comp) program in 2005. It permits Minnesota school districts and teachers' unions to design and collectively bargain teacher-compensation plans that include career ladder/advancement options, job-embedded professional development, teacher evaluation, performance pay, and an alternative salary schedule. The program has led to more comprehensive teacher evaluations in St. Francis and other Minnesota school systems.

For more information: Minnesota Department of Education. mde.q-comp@state.mn.us; http://education.state.mn.us/MDE/Teacher_Support/ QComp/index.html.

Endnotes

[1] Thomas Toch, *In the Name of Excellence* (New York: Oxford University Press, 1991), pp. 138, 167.

[2] *Teaching at Risk: A Call to Action* (New York: The Teaching Commission, 2004).

[3] The Southern Regional Education Board observed in 1988, for example, that "career ladders and other performance-pay incentive programs are the largest educational experiment in the United States today." See Thomas Toch, *In the Name of Excellence*, p. 187.

[4] See Thomas J. Kane and Douglas O. Staiger, "Using Imperfect Information to Identify Effective Teachers" (unpublished paper, School of Public Affairs, University of California-Los Angeles, 2005).

[5] See, for example, David Tyack and Larry Cuban, *Tinkering Toward Utopia: A Century of Public School Reform* (Cambridge, MA: Harvard University Press, 1995).

[6] *State Teacher Policy Yearbook, National Summary, 2007* (Washington, DC: National Council on Teacher Quality, 2007): 92.

[7] "Teacher Rules, Roles and Rights" (Washington, DC: National Council on Teacher Quality, 2007), available online at http://www.nctq.org/cb/. The United States is not the only industrialized country that lacks a commitment to teacher evaluation. Susan Sclafani and Marc Tucker report in a study for the Center on American Progress on teacher compensation in other nations that Germany, a country with a highly regarded educational system, has a long teacher preparation process, demanding entrance standards, and an extended probationary period. But once teachers earn full-time positions, they are evaluated rarely if at all. See Susan Sclafani and Marc Tucker, *Teacher and Principal Compensation: An International Review* (Washington, DC: Center for American Progress, October 2006). p 28 and Gábor Halász, Paulo Santiago, Mats Ekholm, Peter Matthews and Phillip McKenzie, *Attracting, Developing and Retaining Effective Teachers, Country Note: Germany* (Paris, France: Education and Training Policy Division, Directorate for Education, Organisation for Economic Co-operation and Development, September 2004) available online at http://www.oecd.org/dataoecd/32/48/33732207.pdf.

[8] D. Medley and H. Coker, "The Accuracy of Principals' Judgments of Teacher Performance," *Journal of Educational Research* 80, no. 4 (1987): 242.

[9] Brian A. Jacob and Lars Lefgren, "Principals as Agents: Subjective Performance Measurement in Education," Working Paper 11463 (Cambridge, MA: National Bureau of Economic Research, 2005).

[10] *Hiring, Assignment, and Transfer in Chicago Public Schools* (New York: New Teacher Project, August 2007).

[11] John Cronin, Michael Dahlin, Deborah Adkins and G. Gage Kingsbury, *The Proficiency Illusion* (Washington, DC: Thomas B. Fordham Institute and Northwest Evaluation Association, October 2007) available online at http://www.edexcellence.net/doc/The_Proficiency_Illusion.pdf.

[12] Terry Bergner, Julia Steiny, and Jane Armstrong, *Benefits and Lessons Learned from Linking Teacher and Student Data* (Austin TX: National Center for Educational Accountability, December 2007).

[13] *Data Quality Campaign/National Center for Educational Accountability 2007 Survey of State P-12 Data Collection Issues Related to Longitudinal Analysis* (Austin, TX: National Center for Educational Accountability, 2007), available online at http://www.dataqualitycampaign.org/survey_results/. Retrieved January 22, 2008.

[14] Danielson joined the National Institute for Excellence in Teaching as a consultant in 2007.

[15] Mary M. Kennedy, "Recognizing a Good Teacher When You See One" (unpublished paper, Michigan State University, June 2007).

[16] Shujie Liu and Charles Teddlie, "A Follow-up Study on Teacher Evaluation in China: Historical Analysis and Latest Trends," *Journal of Personnel Evaluation in Education* 18, no. 5 (2005): 253–272.

[17] See Kieran M. Killeen, David H. Monk, and Margaret L. Plecki, "School District Spending on Professional Development: Insights from National Data," *Journal of Education Finance* 29 (Summer 2002): 25-50. Numbers updated to 2006 by Education Sector using Department of Labor inflation adjustment calculators.

[18] Mark Wilson, PJ Hallman, Ray Pecheone, and Pamela Moss, "Using Student Achievement Test Scores as Evidence of External Validity for Indicators of Teacher Quality: Connecticut's Beginning Educator Support and Training Program," (unpublished paper, October 2007).

[19] Anthony Milanowski, Steven M. Kimball and B. White, "The Relationship Between Standards-Based Teacher Evaluation Scores and Student Achievement: Replication and Extensions at Three Sites," CPRE-UW Working Paper Series TC-04-01 (Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research, Consortium for Policy Research in Education, 2004): 20.

[20] Robert Pianta, "Spotlight: Classroom Observation, Professional Development and Teacher Quality," *The Evaluation Exchange*, XI, no. 4 (Winter 2005/2006), available online at http://www.gse.harvard.edu/hfrp/eval/issue32/spotlight3.html.

[21] William L. Sanders, James J. Ashton and S. Paul Wright, "Comparison of the Effects of NBPTS-Certified Teachers with Other Teachers on the Rate of Student Academic Progress" (Washington, DC: U.S. Department of Education and National Science Foundation, 2005).

[22] Dan Goldhaber and Emily Anthony, *Can Teacher Quality Be Effectively Assessed* (Washington, DC: Urban Institute, 2004); C. Clotfelter, Helen F. Ladd, J.L. Vigdor, "How and Why Do Teacher Credentials Matter for Student Achievement?" Working Paper 2 (Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2007).

[23] Douglas N. Harris and Tim R. Sass, "The Effects of NBPTS-Certified Teachers on Student Achievement," Working Paper 4 (Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2007).

[24] Douglas N. Harris and Tim R. Sass, "The Effects of NBPTS-Certified Teachers on Student Achievement."

[25] Under a similar "peer assistance and review" program in Rochester, New York, only 80 veteran teachers have been referred to the program since the program's inception nearly two decades ago and only a handful of that group have been fired. See Julia Koppich, "Toward Improving "Teacher Quality: An Evaluation of Peer Assistance and Review in Montgomery County Public Schools, June 2004, p 24.

[26] Jonathan Rochkind, Amber Ott, John Immerwahr, John Doble, and Jean Johnson, *Lessons Learned: New Teachers Talk about Their Jobs, Challenges, and Long-Range Plans: A Report from the National Comprehensive Center for Teacher Quality and Public Agenda* (New York: Public Agenda, 2007).

[27] Ken Futernick, *A Possible Dream: Retaining California Teachers So All Students Can Learn* (Sacramento, CA: Center for Teacher Quality, California State University, 2007).

[28] Dan Goldhaber, Michael DeArmond, and Scott DeBurgomaster, *Teacher Attitudes About Compensation Reform: Implications for Reform Implementation* (Seattle: Center for Reinventing Public Education, 2007): 20.

[29] Lewis C. Solmon, J. Todd White, Donna Cohen and Deborah Woo, *The Effectiveness of the Teacher Advancement Program* (Santa Monica, CA: National Institute for Excellence in Teaching, April 2007): 27.

[30] Teachers earn salary increases rather than bonuses under ProComp. Allan Odden, a co-director of the university-based Consortium for Policy Research in Education and a partner in a compensation consulting company called Teaching Excellence Through Compensation, argues that performance-based bonuses are counterproductive. "If you get rid of base pay, and pay people based on bonuses, people will not want to work in the organization. They don't have a predictable salary to raise a family and pay rent. In an organization like education, you can't have most pay doled out on the basis of bonuses." Personal correspondence with Allan Odden, December 18, 2007.

[31] Thomas Toch, *In the Name of Excellence*, p. 143

[32] Thomas Toch, *In the Name of Excellence*, p 178.

[33] For NEA on peer review, see NEA resolution D-11; for NEA on the use of student test scores to evaluation teachers, see NEA resolution D-20.

[34] NEA resolution F-9.

[35] So is the Chinese education system. Researchers Shujie Liu of the University of Southern Mississippi and Charles Teddlie of Louisiana State University report in a study of Chinese teacher evaluation practices that the Chinese education system is combining 360-degree evaluations with classroom observations and students achievement. See Shujie Liu and Charles Teddlie, "A Follow-up Study on Teacher Evaluation in China: Historical Analysis and Latest Trends."

[36] Fred M. Newmann, Anthony S. Bryk, and Jenny Nagaoka, *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* (Chicago: Consortium on School Reform, 2001).

[37] Robert Gordon, Tom Kane and Doug Staiger make this point in a Brookings Institution report. Robert Gordon, Thomas J. Kane and Douglas O. Staiger, "Identifying Effective Teachers Using Performance on the Job," A Hamilton Project Discussion Paper (Washington, DC: The Brookings Institution, April, 2006).

Bibliography

Adolescence and Young Adulthood/Mathematics Standards (Arlington, VA: National Board for Professional Teaching Standards, 2001).

Archibald, Sarah, "How Well Do Standards-Based Teacher Evaluation Scores Identify High-Quality Teachers? A Multilevel, Longitudinal Analysis of One District." PhD diss., University of Wisconsin-Madison, 2007.

Borman, Geoffrey D. and Steven M. Kimball, "Teacher Quality and Educational Equality: Do Teachers With Higher Standards-Based Evaluation Ratings Close Student Achievement Gaps?" *The Elementary School Journal* 106, no. 1 (2005): 3–20.

Caldwell, Tanya, "Board Vows to Cover Teachers' Merit Pay," *Orlando Sentinel*, August 14, 2007.

Campbell, Donald J., Kathleen M. Campbell and Ho-Beng Chia, "Merit Pay, Performance Appraisal, and Individual Motivation: An Analysis and Alternative," *Human Resource Management* 37, no. 2 (Summer 1998): 131–146.

Carey, Kevin, "The Real Value of Teachers," *Thinking K–16* Vol. 8, no. 1 (Winter 2004), (Washington, DC: The Education Trust).

Cavaluzzo, Linda C., *Is National Board Certification an Effective Signal of Teacher Quality?* (Alexandria, VA: CNA Corporation, 2004).

Cincinnati Public Schools, "Teacher Evaluation," http://www.cps-k12.org/employment/tchreval/tchreval.htm.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor, "How and Why Do Teacher Credentials Matter for Student Achievement?" Working Paper 2 (Washington DC: National Center for Analysis of Longitudinal Data in Education Research, 2007).

Community Training and Assistance Center, *Catalyst for Change: Pay for Performance in Denver, Final Report* (Boston: Community Training and Assistance Center, January 2004).

Connecticut State Department of Education, *A Guide to the BEST Program for Beginning Teachers, 2006–2007* (Hartford, CT: Connecticut State Department of Education).

Connecticut State Department of Education, *Portfolio Performance Results, Five Year Report, 1999–2004*. (Hartford, CT: Connecticut State Department of Education, August 2005).

Creating a Successful Performance Compensation System for Educators (Washington DC: National Institute for Excellence in Teaching, July 2007).

Danielson, Charlotte, *Enhancing Professional Practice: A Framework for Teaching*, 2nd ed. (Alexandria, VA: Association for Supervision and Curriculum Development, 2007).

Danielson Charlotte and Thomas L. McGreal, *Teacher Evaluation to Enhance Professional Practice* (Alexandria, VA: Association for Supervision and Curriculum Development, 2007).

Darling-Hammond, Linda and Cynthia D. Prince, *Executive Summary: Strengthening Teacher Quality in High-Need Schools: Policy and Practice* (Washington, DC: Council of Chief State School Officers, 2007).

Figlio, David N. and Lawrence W. Kenny, "Individual Teacher Incentives and Student Performance," *Journal of Public Economics* 91, no. 5–6 (June 2007): 901–914.

Goldhaber, Dan, *Everybody's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?* CALDER Working Paper 9 (Washington, DC: Urban Institute, April 2007).

Goldhaber, Dan and Emily Anthony, *Can Teacher Quality Be Effectively Assessed?* (Washington, DC: Urban Institute, 2004).

Goldhaber, Dan, Michael DeArmond, Scott DeBurgomaster, *Teacher Attitudes About Compensation Reform: Implications for Reform Implementation* (Seattle: Center for Reinventing Public Education, 2007).

Goldhaber, Dan, Michael DeArmond, Albert Liu and Dan Player, *Returns to Skill and Teacher Wage Premiums: What Can We Learn by Comparing the Teacher and Private Sector Labor Markets?* (Seattle: Center for Reinventing Public Education, 2007).

Gonring, Phil, Paul Teske, and Brad Jupp, *Pay-for-Performance Teacher Compensation* (Cambridge, MA: Harvard Education Press, 2007).

Gordon, Robert, Thomas J. Kane and Douglas O. Staiger, "Identifying Effective Teachers Using Performance on the Job," A Hamilton Project Discussion Paper (Washington, DC: The Brookings Institution, April, 2006).

Halverson, Richard, Carolyn Kelly and Steven M. Kimball, "Implementing Teacher Evaluation Systems: How Principals Make Sense of Complex Artifacts to Shape Local Instructional Practice," in *Research and Theory in Educational Administration Volume 3*, eds. W. Hay and C. Miskel (Greenwich, CT: George F. Johnson, 2003): 153–188.

Hanushek, Eric A., "The Tradeoff Between Child Quantity and Quality," *Journal of Political Economy* 100, no. 1 (1992): 84–117.

Harris, Douglas N. and Tim R. Sass, "The Effects of NBPTS-Certified Teachers on Student Achievement," Working Paper 4 (Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2007).

Heneman, H.G. III and Anthony Milanowski, "Alignment of Human Resource Practices and Teacher Performance Competency," *Peabody Journal of Education* 79, no. 4 (2004): 108–125.

Heneman, H.G. III, Anthony Milanowski, Steven M. Kimball, and Allan Odden, *Standards-Based Teacher Evaluation as a Foundation for Knowledge- and Skill-Based Pay*. CPRE Policy Brief RB-45. (Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education, May 2006).

Hershberg, Ted, "Value-Added Assessment and Systemic Reform: A Response to America's Human Capital Development Challenge," (paper prepared for the Aspen Institute's Congressional Institute, Cancun, Mexico, February 22–27, 2005).

Hobbs, Erika, "Merit Pay for Teachers Reveals Sway of Affluence," *Orlando Sentinel*, September 9, 2007.

Holtzapple, Elizabeth, "Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System," *Journal of Personnel Evaluation in Education* 17, no. 3 (2003): 207–219.

Jacob, Brian A. and Lars Lefgren, "Principals as Agents: Subjective Performance Measurement in Education," Working Paper 11463 (Cambridge, MA: National Bureau of Economic Research, 2005).

Kane, Thomas J., J.E. Rockoff, and Douglas O. Staiger, *What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City* (Cambridge, MA: National Bureau of Economic Research, 2006).

Kanstoroom, Marci and Chester E. Finn, Jr., eds., *Better Teachers, Better Schools* (Washington, DC: Thomas B. Fordham Institute, July 1999).

Kellor, Eileen M., *Performance-Based Licensure in Connecticut* (Madison, WI: Consortium for Policy Research in Education, University of Wisconsin-Madison, 2002).

Kennedy, Mary M., "Monitoring and Assessing Teacher Quality," (paper presented at the annual meeting of the American Educational Research Association, Chicago, April 2007).

Kennedy, Mary M., "Recognizing a Good Teacher When You See One," (unpublished paper, Michigan State University, June 2007).

Killeen, Kieran M., David H. Monk, and Margaret L. Plecki, "School District Spending on Professional Development: Insights from National Data," *Journal of Education Finance* 29 (Summer 2002): 25–50.

Kimball, Steven M., "Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems," *Journal of Personal Evaluation in Education* 16, no. 4 (2002): 241–268.

Koppich, Julia, "Toward Improving Teacher Quality: An Evaluation and Review in Montgomery County Public Schools," available online at http://www.mcps.k12.md.us/departments/development/documents/pgs/PAR_report_final.doc

Liu, Shujie and Charles Teddlie, "A Follow-up Study on Teacher Evaluation in China: Historical Analysis and Latest Trends," *Journal of Personnel Evaluation in Education* 18, no. 5 (2005): 253–272.

McCaffrey, Daniel F., J.R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability* (Santa Monica, CA: Rand, 2003).

Milanowski, Anthony, "Relationships Among Dimension Scores of Standards-Based Teacher Evaluation Systems and the Stability of Evaluation Score/Student Achievement Relationships Over Time," CPRE-UW Working Paper Series TC-04-02 (Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research, Consortium for Policy Research in Education, San Diego, CA, 2004).

Milanowski, Anthony, "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati," *Peabody Journal of Education* 79, no. 4 (2004): 33–53.

Milanowski, Anthony and H.G. Heneman III, "Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study," *Journal of Personnel Evaluation in Education* 15, no. 3 (2001): 193–212.

Milanowski, Anthony and Steven M. Kimball, "The Framework-Based Teacher-Evaluation Systems in Cincinnati and Washoe," CPRE-UW Working Paper Series TC-03-07 (Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education, April 2003).

Milanowski, Anthony, Steven M. Kimball and Allan Odden, "Teacher Accountability Measures and Links to Learning," in *Measuring School Performance and Efficiency: Implications for Practice and Research*, eds. Leanna Stiefel, Amy Ellen Schwartz, Ross Rubenstein and J. Zabel (Yearbook of the American Education Finance Association, 2005): 137–161.

Milanowski, Anthony, Steven M. Kimball and B. White, "The Relationship Between Standards-Based Teacher Evaluation Scores and Student Achievement: Replication and Extensions at Three Sites," CPRE-UW Working Paper Series TC-04-01 (Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research, Consortium for Policy Research in Education, 2004).

Milanowski, Anthony, Allen Odden, Brad White, E. Kellor, H.G. Heneman III, James Allen and Kimberly Mack, "Final Report on the Evaluation of the 2000–2001 Implementation of the Cincinnati Federation of Teachers/Cincinnati Public Schools Teacher Evaluation System," Working Paper TC-01-3 (Madison, WI: University of Wisconsin, Wisconsin Center for Education Research, Consortium for Policy Research in Education, July 2001).

The New Teacher Project, *Hiring, Assignment, and Transfer in Chicago Public Schools* (New York: The New Teacher Project, August 2007).

Odden, Allen, "Lessons Learned About Standards-Based Teacher Evaluation Systems," *Peabody Journal of Education* 79, no. 4 (2004): 126–137.

Odden, Allan and Marc Wallace, *How to Achieve World Class Teacher Compensation* (Freeload Press, 2008).

Ovando, Martha N., "Building Instructional Leaders' Capacity to Deliver Constructive Feedback to Teachers," *Journal of Personal Evaluation in Education* 18, no. 3 (2004): 171–184.

Pecheone, Raymond L. and Ruth R. Chung, "Evidence in Teacher Education the Performance Assessment for California Teachers (PACT)," *Journal of Teacher Education* 57, no. 1 (January/February 2006): 22–36.

Peterson, K.D., *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices* (Thousand Oaks, CA: Corwin Press, 2000).

Pianta, Robert C., "Standardized Observation and Professional Development: A Focus on Individualized Implementation and Practices," in *Critical Issues in Early Childhood Professional Development*, eds. M. Zaslow and I. Martinez-Beck (Baltimore: Brookes Publishing, 2005), 231–254.

Podgursky, Michael J. and Matthew G. Springer, *Teacher Performance Pay: A Review* (Nashville, TN: National Center on Performance Incentives, 2006).

A Research Guide on National Board Certification of Teachers (Arlington, VA: National Board for Professional Teaching Standards, 2007).

Sanders, William L., James J. Ashton and S. Paul Wright, "Comparison of the Effects of NBPTS-Certified Teachers with Other Teachers on the Rate of Student Academic Progress" (Washington, DC: U.S. Department of Education and National Science Foundation, 2005).

Schemo, Diana Jean, "When Students' Gains Help Teachers' Bottom Line," *The New York Times*, May 9, 2004.

Sclafani, Susan and Marc S. Tucker, *Teacher and Principal Compensation: An International Review* (Washington, DC: Center for American Progress, October 2006).

State Teacher Policy Yearbook Progress on Teacher Quality 2005 (Washington, DC: National Council on Teacher Quality, 2005).

State Teacher Policy Yearbook Progress on Teacher Quality 2007 (Washington, DC: National Council on Teacher Quality, 2007).

Takakura, Sho and Yumika Ono, "Restructuring Teacher Evaluation in Japan: Recent Developments in Personnel Management System," (paper presented at the annual meeting of the Japan-U.S. Teacher Education Consortium, Tacoma, WA, August 2001).

The Teaching Commission, *Teaching at Risk: A Call to Action* (New York: The Teaching Commission, 2004).

The Teaching Commission, *Teaching at Risk, Progress and Potholes* (New York: The Teaching Commission, 2006).

Toch, Thomas, *In the Name of Excellence* (New York: Oxford University Press, 1991).

Wallace, Marc J. Jr., *[School or District] Teacher Evaluation System DRAFT January 2006 Portfolio Development Handbook* (Lake Bluff, IL: Teacher Excellence Through Compensation, 2006).

Wilson, Mark, PJ Hallman, Ray Pecheone, and Pamela Moss, "Using Student Achievement Test Scores as Evidence of External Validity for Indicators of Teacher Quality: Connecticut's Beginning Educator Support and Training Program," (unpublished paper, October 2007).

Wise, Arthur E., Linda Darling-Hammond, Milbrey W. McLaughlin and Harriet T. Bernstein, *Teacher Evaluation A Study of Effective Practices* (Santa Monica, CA: Rand Corporation, 1984).

White, Brad, "The Relationship Between Teacher Evaluation Scores and Student Achievement: Evidence from Coventry," RI. CPRE-UW Working Paper Series TC-04-04 (Madison, WI: University of Wisconsin-Madison, Wisconsin Center Education Research, Consortium for Policy Research in Education, San Diego, CA, 2004).

Youngs, Peter, "District Induction Policy and New Teachers' Experiences: An Examination of Local Policy Implementation in Connecticut," *Teachers College Record* 109, no 3 (2007): 797–837.

Youngs, Peter, "How Elementary Principals' Beliefs and Actions Influence New Teachers' Experiences," *Education Administration Quarterly* 43, no. 1 (2007): 101–137.

Youngs, Peter, "State and District Policy Related to Mentoring and New Teacher Induction in Connecticut," (paper prepared for the National Commission on Teaching and America's Future, December 2002).